

# METHODOLOGICAL ISSUES WITH STUDENT EVALUATION OF TEACHING EFFECTIVENESS (SETE) - PART 2

Kevin Brewer

Orsett Technical Reports, Series A, No.3

ISBN: 978-0-9540761-6-0

PUBLISHED BY  
Orsett Psychological Services,  
PO Box 179,  
Grays,  
Essex  
RM16 3EW  
UK

COPYRIGHT  
Kevin Brewer 2002

## COPYRIGHT NOTICE

All rights reserved. Apart from any use for the purposes of research or private study, or criticism or review, this publication may not be reproduced, stored or transmitted in any form or by any means, without prior permission in writing of the publishers. In the case of reprographic reproduction only in accordance with the terms of the licences issued by the Copyright Licensing Agency in the UK, or in accordance with the terms of licences issued by the appropriate organization outside the UK.

## **General Series Introduction**

Orsett Technical Reports are designed to allow the exploration of specific topics in detail. Series A contains four reports on different aspects of the student evaluation of teaching effectiveness (SETE) or students' ratings of instruction (SRI). This is the rating of lecturers and teachers by their students.

### **REPORT No.1 <sup>1</sup>**

This report is a literature review of the studies into SETE and SRI, mostly from the USA. The aim is to outline what students see as the "ideal lecturer". Much of the material comes from the prolific work of Kenneth Feldman.

### **REPORT No.2 <sup>2</sup>**

This report addresses the issue of the accuracy of students' ratings of their instructors. Is it an accurate picture of their teaching effectiveness or the personal feelings of the students? The issues of reliability, generalisability, and validity of the ratings, along with rating errors, are discussed.

### **REPORT No.3 <sup>3</sup>**

**Report no.3 takes many of the technical issues raised in report no.2 further. In particular, the potential biases to SETE and SRI.**

### **REPORT No.4 <sup>4</sup>**

This report gives details of the construction of the Birmingham Overseas Student Teaching Evaluation Questionnaire (BOSTEQ). The aim is to produce a rating instrument specifically to be used by overseas students.

The research is part of an MSc degree at the University of Aston <sup>5</sup>.

---

<sup>1</sup> Brewer, K (2002a) Student evaluation of teaching effectiveness: an introduction, Orsett Technical Reports, Series A, No.1, Orsett Psychological Services: Orsett, Essex.

<sup>2</sup> Brewer, K (2002b) Student evaluation of teaching effectiveness: methodological issues - part 1, Orsett Technical Reports, Series A, No.2, Orsett Psychological Services: Orsett, Essex.

<sup>3</sup> Brewer, K (2002c) Methodological issues with student evaluation of teaching effectiveness (SETE) - part 2, Orsett Technical Reports, Series A, No.3, Orsett Psychological Services: Orsett, Essex.

<sup>4</sup> Brewer, K (2002d) Construction of Birmingham Overseas Students Teaching Evaluation Questionnaire (BOSTEQ), Orsett Technical Reports, Series A, No.4, Orsett Psychological Services: Orsett, Essex.

<sup>5</sup> Brewer, K (1993) Overseas Students Evaluation of Teaching Effectiveness, Unpublished MSc thesis, University of Aston: Birmingham, UK.

# **CONTENTS**

|                                                                                                   | Page No. |
|---------------------------------------------------------------------------------------------------|----------|
| CAN STUDENTS JUDGE GOOD LECTURERS?                                                                | 4        |
| FACTORS AFFECTING STUDENT EVALUATION OF TEACHING                                                  | 4        |
| 1. Actual/expected grades                                                                         | 5        |
| 2. Class size                                                                                     | 6        |
| 3. Prior subject interest                                                                         | 7        |
| 4. Instructor rank/experience                                                                     | 7        |
| 5. Sex of instructor/student                                                                      | 8        |
| 6. Instructor expressiveness                                                                      | 9        |
| 7. Characteristics of course                                                                      | 9        |
| 8. Student's personality                                                                          | 10       |
| 9. Reason for rating                                                                              | 11       |
| 10. Administration of ratings                                                                     | 11       |
| CAN STUDENT RATINGS OF TEACHERS BE TRUSTED?                                                       | 13       |
| RELIABILITY OF STUDENT RATINGS                                                                    | 13       |
| VALIDITY OF STUDENT RATINGS                                                                       | 16       |
| Objective validation: criterion validation                                                        | 16       |
| Construct validation                                                                              | 19       |
| REFERENCES                                                                                        | 23       |
| APPENDIX 1 - Other less important potential biases<br>of student ratings of instruction           | 29       |
| APPENDIX 2 - Class size and individual<br>characteristics of teaching effectiveness               | 34       |
| APPENDIX 3 - Correlation of individual items with<br>overall evaluation of teaching effectiveness | 36       |

## **CAN STUDENTS JUDGE GOOD LECTURERS?**

Marris (1964) says students "are still the best judges of a course of lectures, if only because they are generally the only people who listen to them" (quoted by Cooper and Foy 1967 p182). Similarly, Riley et al (1950) conclude that "the students' construct of 'good teaching' is closely relevant to the effectiveness of a teacher in reaching the students" (quoted in Flood Page 1974 p29).

Yet not everybody would agree. Bryant (1967) is scathing: "Most undergraduate students, after all, are not yet fully mature. They do not understand what they can get, should get, or will need from a college education" (quoted in Flood Page 1974 p25). He suggests that students evaluate courses based on what is "fun" or "dull", not what is learned.

Cooper and Foy's (1967) checklist of the ideal lecturer was objected to on the basis that "student opinion is worthless"; students seek different characteristics at different times/classes; and the characteristics observed in the lecturer are based on the interaction with that group (Foy 1969).

So, in summary, the main arguments against student ratings of teaching are:

i) Students' decisions/evaluations are influenced by factors other than just the lecture (this is the issue of whether student ratings of instruction are biased).

ii) Students do not know what is a good lecture and teaching (this is the question of the validity of student ratings of instruction).

iii) Students change their minds over time (this is concerned with the reliability of student ratings of instruction).

These three issues are at the crux of whether student evaluation of teaching effectiveness can be trusted.

## **FACTORS AFFECTING STUDENT EVALUATION OF TEACHING**

Because of the varied reasons for using student evaluation of teaching, the timing of the administration of the instrument varies. Thus extraneous variables become important. An expectation or evaluation is open to many influences. There is a fear that the instrument, especially one administered well into a course, will

measure something other than simply the student's feeling about the lecturer/course.

And this undermines the usefulness of the instruments, say those with this fear. Marsh (1984), however, believes there is a "witch hunt for potential biases" (p730).

Dunkin and Barnes (1986) are not afraid: "The usefulness of student evaluation does not depend on their being free of such influences, so much as the ability to take account of them" (p769).

Here are some of the main factors that could influence the students' ratings of the lecturer/course:

1. actual/expected grades;
2. class size;
3. prior subject interest;
4. instructor rank/experience;
5. sex of instructor/student;
6. instructor expressiveness;
7. characteristics of the course;
8. student's personality;
9. reasons for rating;
10. administration of ratings.

## 1. ACTUAL/EXPECTED GRADES.

Generally classes expecting or possessing higher grades give higher ratings. This is sometimes known as the "grading bias hypothesis". A number of studies support this (see Arubayi 1987), yet others also contradict it (eg: Bendig 1953).

Brown (1976) uses multiple regression analysis to conclude that grades do bias student ratings. Grades only accounted for 9% of the variance, but this is more than the other variables (eg: class size, course level).

However, the findings are not always consistent. Cohen (1981) embarked on a meta-analysis<sup>6</sup> of 41 studies on this question, and was able to reject the null hypothesis of no relationship between course rating and grades. But within the 41 studies were variations in findings. Marsh (1984) discusses possible reasons for the findings.

But the positive relationship does not always exist across all situations. For example, Anikeef (1953) found a stronger relationship between expected grade and the lower the level of the class. Also there are differences across all aspects of teaching. Echandia (1964) looked at accounting students: those who received higher grades

---

<sup>6</sup> See Glass 1974; 1978; McCallum 1984 for more details on meta-analysis.

rated the lecturer as better organised and as having clearer presentation than those with lower grades, but there was no difference on the lecturer's ability to motivate students.

Feldman (1976a) remains undecided after reviewing over 200 correlations; concluding that "it cannot be said that grades tend to bias evaluation. But neither can it be concluded that they do not" (p100).

## 2. CLASS SIZE.

Flood Page (1974) feels that "either class size makes no difference, or that larger classes tend to give worse ratings" (p58).

Feldman (1978) reviewed 50 studies of which 17 showed no significant relationship between class size and student ratings. The other 33 showed either a small negative correlation, or a curvilinear relationship (ie: higher ratings to teachers of small and large classes compared to medium sized classes). The author attempts to explain the curvilinear relationship on the basis that increased resources are given to larger classes, or particular instructors are chosen who can teach large classes well, or instructors see the large class as a challenge and put more effort into preparation.

In a further review, Feldman (1984) found two studies with significant positive correlations, 22 studies with no relationship, 22 with a small negative relationship, and 8 showing a curvilinear relationship. Ignoring the curvilinear relationship, the average correlation was only  $r = -.09$ .

Feldman then compared the studies showing the relationship between individual characteristics of teaching and class size. Most characteristics showed no relationship, except a negative correlation of class size with presentation of subject matter, and communication.

Feldman concludes that "class size has been found to be related more frequently and with greater strength to those instructional dimensions involving teachers' interactions and interrelations with students" (1984 p77).

Frey (1978) testing two dimensions of student rating ("skill" and "rapport") against class size found a strong, negative relationship between class size and ratings of "rapport", while the "skill" factor showed a weak, positive relationship. This agrees with Costin et al's (1971) statement that the relationship "may vary according to the particular aspect of teaching performance that the student is asked to rate" (p521).

### 3. PRIOR SUBJECT INTEREST.

Marsh and Cooper (1981) looked at the correlation between the student rating of the instructor, and the students' prior subject interest, using 511 undergraduates in Southern California. The correlation was 0.2 ( $p < 0.01$ ) for overall rating, but varied for different dimensions of teaching.

Marsh (1982a) examines 16 student/course instructor characteristics, and found that prior subject interest was the variable with the largest impact on ratings. But concludes here that "lecturers actually are more effective at teaching when working with motivated students, and that this more effective teaching is accurately reflected in the student ratings" (p85).

Prior subject interest is a bias, but not specifically to student ratings of instruction; for example, students with high prior subject interest usually do well in course examinations also.

### 4. INSTRUCTOR RANK/EXPERIENCE.

Here there are mixed findings, but probably little effect (Marsh 1985). Frey (1978) reports the concern that younger instructors will get higher ratings because students identify more closely with them. Some evidence supports this (eg: Clark and Keller 1954; Guthrie 1949, 1954).

However, Arubayi (1987) reports studies showing that professors receive higher ratings than lecturers (eg: Downie 1952; Gage 1961).

Frey provides an answer to this contradiction at the Northwestern University, Illinois using his two dimensions of student ratings ("rapport" and "skill"). The ratings on the "rapport" factor decreased steadily with rank/age, while the "skill" factor showed the opposite trend.

Feldman (1983), in another of his extensive reviews, compared a number of studies under three headings - academic rank, age, and instructional experience. Table 1 shows a summary of the studies found by Feldman, and the type of correlations these studies found.

These three distinctions of rank, age, and instructional experience help to account for the mixed findings. The relationship of academic rank to teaching effectiveness evaluation has more significant positive correlations suggesting that the higher the rank, the more positive the student rating of the instruction. While age has no significant positive correlations,

| Type of study          | Studies finding significant positive correlation | Studies finding significant negative correlation | Studies finding no correlation | Studies finding other patterns |
|------------------------|--------------------------------------------------|--------------------------------------------------|--------------------------------|--------------------------------|
| ACADEMIC RANK          | 10                                               | 1                                                | 21                             | 1                              |
| AGE                    | 0                                                | 6                                                | 6                              | 0                              |
| INSTRUCTION EXPERIENCE | 2                                                | 5                                                | 8                              | 1                              |

Table 1 - showing the number of studies found by Feldman (1983) showing the different relationships between seniority and teaching effectiveness.

suggesting that older lecturers are not rated more positively than younger lecturers.

## 5. SEX OF INSTRUCTOR/STUDENT.

### SEX OF THE STUDENT:

Doyle and Whitely (1974) felt it was generally unrelated or trivial. Arubayi (1987) reports the conclusion from findings that female students rate more favourably than male students; and they rate female lecturers more highly than male lecturers. Aleamoni and Thomas (1977) report no relationship between sex of rater and rating of faculty.

### SEX OF THE INSTRUCTOR:

Feldman (1992) reviewed 14 experiments producing 485 analyses, and found that for overall evaluation, there was no difference in the ratings based on gender of the lecturer. Then he examined the individual characteristics of teaching. Again, generally no difference, but if there was a difference, male teachers received higher ratings.

In the second half of the article, Feldman (1993) reviewed classroom studies finding no general difference, but this time, if there was a difference it favoured women. The average correlation was only  $r = +.02$  between the sex of the instructor and overall evaluation of teaching.

Martin (1984) found that instructors who fitted social stereotypes received better evaluations.

Developing this idea, D'Agostino and Dill (1988) noted that behaviours classed as friendliness towards



students produced higher ratings for female instructors, but not for male. But overall, male professors were rated as more effective than female professors. The authors conclude that "male and female instructors will earn equal SRI (student ratings of instruction) for equal professional work only if the women also display stereotypical feminine behaviour" (p344).

## 6. INSTRUCTOR EXPRESSIVENESS.

Instructor's expressiveness has sometimes been studied as the "Dr. Fox paradigm": in that students give high ratings to entertaining lecturers even though the content is nonsense. Based on the original work of Naftulin, Ware and Donnelly (1973), who introduced an actor as Dr. Myron. L. Fox to give a lecture to a group of educators and mental health professionals. He was entertaining, but spoke deliberate nonsense. Naftulin et al suggested that the lecturer's expressiveness can "seduce" students into believing they have learned something.

Abrami, Leventhal and Perry (1982) compiled a meta-analysis of the studies on the "Dr.Fox paradigm", finding inconsistencies. They conclude that "instructor expressiveness had a substantial impact on student ratings but a small impact on student achievement" (p446), while lecture content had the opposite relationship.

The methodology of the original experiment by Naftulin et al has been criticised heavily by Frey (1978) and Marsh (1984).

## 7. CHARACTERISTICS OF THE COURSE.

### WORKLOAD:

Marsh (1984) quotes his own earlier research (1982b), where two courses given by the instructor were compared. The course perceived as having the heavier workload or being more difficult was rated higher. However, Marsh does not believe this causes a bias to student ratings.

### REASON FOR COURSE:

Research has compared optional against compulsory courses, with teachers of the latter being rated lower sometimes. While those students taking the subject as a

major tend to give more positive ratings of the lecturer than non-majors (Feldman 1978).

#### CLASS LEVEL:

Feldman (1977) notes an inconsistency in results of studies on class level and ratings. He suggests it is due to failure to take account of other factors. Marsh and Overall (1981) looked at the contribution of course level (ie: undergraduate or postgraduate), and of course type in determining evaluations of teaching. The former was not statistically significant, while the latter was, but accounted for no more than 2-3% of variance on ANOVA analysis. The effect of a specific instructor accounted for five to ten times as much variance on the same analysis.

#### 8. STUDENT'S PERSONALITY.

Rezler (1965) administered the Purdue Rating Scale for Instruction (Remmers 1960), and the Edwards Personal Preference Schedule (which assesses student needs). They found several significant correlations:

- Male students with high needs for "nurturance", heterosexual relations, exhibitionism, and dominance rated male teachers higher.

- Female students with high needs for "succorance", heterosexual relationships, and exhibitionism rated all teachers lower (quoted in Flood Page 1974).

Smithers (1970b), working at the University of Bradford, has looked at students' scores on the Eysenck Personality Inventory (Eysenck and Eysenck 1964) and Rokeach's dogmatism scale (Rokeach 1960), and their expectations of the lecturers. Significant differences ( $p < 0.05$ ) were found on nine of the 50 items.

Extraverts expected the lecturer to be "entertaining and confident" compared to introverts; low scorers on neuroticism are less concerned about "speed of lecture" and "lecturer setting a standard" compared to high scorers. Neurotic introverts are less concerned about the "lecturer taking own line on controversial issues", and want less "use of non-textbook material" compared to other students.

High dogmatism scorers have significantly higher expectations on four items compared to low scorers: "keeps to point"; "thoroughly prepares for lecture"; "provides duplicated notes of lecture"; and "organises blackboard work clearly".

Other studies have found differences in student ratings based on differences in authoritarianism (Freehill 1967); and general personality profile (eg: Rees 1969; Yonge and Sassenrath 1968). Flood Page (1974) concludes that "there is some kind of slight effect, but not one of any practical importance" (p50).

Feldman (1977) feels it is difficult to generalise:

Direction and content differences seem dependent on the nature of the rating items, the specific personality or related characteristics measured, differences in experiences and other attributes of the student, and the particulars of the courses and teachers (p244).

## 9. REASONS FOR RATING.

Ratings being used for promotion purposes are generally higher. Tetenbaum (1977) asked 414 students to evaluate their instructors, and they were divided into three conditions. The difference being the supposed purpose of the ratings - for promotion purposes; to improve quality of teaching; or to aid future course selection. The three conditions produced different means, and then slight variations in the factor analysis.

Feldman (1979), in an extensive review, concludes that the ratings are higher for "official" purposes (eg: promotion), but the studies are limited, so caution is needed.

## 10. ADMINISTRATION OF RATINGS.

### ANONYMOUS VS IDENTIFIED:

It is generally felt that identified ratings are higher, but Feldman (1979) emphasises the context in which students identify themselves. For example, when students were asked to identify themselves to explain their evaluations later, the ratings were always higher, than when identified but "only for research purposes" (Sharon and Bartlett 1969).

### WHO ADMINISTERS EVALUATION QUESTIONNAIRE:

Kirchner (1967) found significant differences in ratings, between when the instructor or neutral observer administers the evaluation session. Presence of the instructor while being evaluated leads to higher ratings.

Other factors include the instructor's demeanour during the rating if they are present; presence of the instructor's colleagues (produces higher ratings); and rapport between ratee and rater (Doyle 1983).

#### WHEN EVALUATION QUESTIONNAIRE ADMINISTERED:

This not important if "(1) the students are asked to rate typical performance; (2) they have had sufficient opportunity to observe the instructor; (3) the evaluations do not take place at the same time as special events like holidays, perhaps, or examinations that might influence the data" (Doyle 1975 p78).

Frey (1976) feels that ratings administered during final exams are generally lower than those administered during term.

Feldman (1979) points out that the few studies that have compared the timing of the evaluation have not found any differences.

#### RATING FORMAT:

Feldman (1979) includes three variables related to the format of the rating instrument that could influence the results:

i) The instructions given to the students on how to fill in the rating form.

ii) The items used ("stimulus variables").

iii) The response options available. Follman et al(1974) offered three groups of students different responses - "degree of agreement" with statement; degree to which improvement needed in characteristic given; and ordered categories (eg: "excellent", "average"). The first two produced higher ratings (non-significant though).

Feldman (1979) details other variations in rating formats that can influence the level of student ratings:

- that lead to higher ratings:
  - the use of "degree of agreement" rather than disagreement; dropping unfavourable response items but keeping the same number of items; and positive phrasing of the "stem".
- that seem to have no effect:
  - amount of information about the trait being assessed; offering only positive/neutral response items;

varying the order of response categories; using negative numbers as response items; or the type of person to imagine (eg: "ideal teacher" or "best teacher you have had").

There are a number of other possible biases to the student rating of instruction, but these are seen as less important in the literature. Details of these can be found in the Appendix 1.

## **CAN STUDENT RATINGS OF TEACHERS BE TRUSTED?**

Opinions vary over the faith to put in student ratings, particularly because of the infinite number of background variables that could bias the ratings.

An interesting study is reported by Miklich (1969) from the University of Hawaii. He compares two groups he had to teach - one he knew well, the other he was teaching for the first time. For the latter he took pains to explain the examinations. The student ratings from the two groups showed a significant difference: "Fairness of Grading" was rated higher by the new group. This seems to suggest that the students were responding to the teacher's behaviour.

Marsh, who has written extensively in this area, believes in the system of student ratings, as long as expectations are not too high. Most studies, he reports, have found a correlation of 0.30 or less between student ratings and particular variables (Marsh 1984).

Marsh (1982a) concludes "that none of the suspected biases to student ratings seems actually to have much impact" (p87).

Furthermore, Dunkin and Barnes (1986) finish their literature review reasonably confident that students can perceive and rate their teaching. So background variables do not invalidate the idea that students can tell what is a good lecture. But whether student ratings are valid and reliable, which are important issues before we can trust them, will be reviewed next.

## **RELIABILITY OF STUDENT RATINGS**

Do students change their minds over time, or maybe vary in the ratings from class to class due to inconsistency?

Here reliability refers to the fact that the ratings will measure the same score every time, ie: the same lecturer producing the same quality lecture on two occasions will receive the same rating by the same

student.

Doyle (1975) lists the sources of reliability errors:

- i) Computational error - eg: putting the wrong instructor's name on ratings summary.
- ii) Rater's task - ie: problem with nature of the questions used.
- iii) Environment - physical or social environment.
- iv) Rater - lacks motivation or memory problems.
  - Halo effect: overall impression influences specific rating items.
  - Leniency error: tendency to rate higher when known that ratings being used for promotion purposes.
  - Central tendency: inclination for mid-point on scale.
  - Proximity error: rate adjacent items similarly.
  - Contrast error: projection of own deficiencies on to ratee.
  - Logical error: rating traits that "ought" to go together.

The first study of reliability came from Guthrie (1927). 285 psychology students ranked lecturers at the University of Washington, and then again two weeks later. A correlation of  $r = 0.89$  was found.

Foy (1969) followed up his study with Cooper (Cooper and Foy 1967), due to objections about the original findings on an ideal lecturer. A different group of students used the same check-list as the first study, and there was a correlation of 0.93 between the two ratings (1 in 2000 possibility of a chance correlation as high as that). This seems the most straightforward evidence of the reliability of an instrument. Arubayi (1987) reviews a number of studies, "from what is available in the literature it appears that student ratings are reasonably reliable" (p269).

The reliability of individual instruments obviously is an important requirement before general use. Bradbury and Ramsden (1975) detail a reliability retest of the North East London Polytechnic student feedback questionnaire between 7 to 14 days after the original use. The reliability coefficient was 0.77 or above. Certainly for the well-established questionnaires, reliability coefficients are as expected - eg: Marsh (1982a) testing the reliability of SEEQ finds correlations of between 0.74 - 0.90 using intra-class correlations (random half of class correlated to other half), and coefficient alpha between 0.88 - 0.97.

Overall and Marsh (1980) found a significant correlation between ratings in the final year of a course, and one year afterwards; using both class-average responses (median  $r = 0.83$ ) and individual student responses (median  $r = 0.58$ ).

Another way of looking at reliability could be to compare two groups cross-sectionally. Drucker and Remmers (1951) compared current undergraduates with alumni of 10 years, for ranking of ideal lecturer, using the Purdue Rating Scale for Instruction (Remmers 1960). Of ten items, there was agreement on seven, including the first four: "presentation of subject matter", "interest in subject", "stimulating intellectual curiosity", and "liberal and progressive attitude".

Centra (1974) adapted this study to look at overall assessment of teaching between current students and alumni (of five years). There was a significant correlation ( $r = .75$ ) between the two groups on the rating of "best" and "worst" lecturers in the department.

So students' idea of what constitutes a good teacher remains similar as they grow older.

Braskamp et al (1985) make a number of generalisations about the reliability of SRI:

1. Student agreement on global ratings are sufficiently high if class greater than 15 students <sup>7</sup>.
2. Students are consistent in their global ratings of the same instructor at different times in the course <sup>8</sup>.
3. An instructor's overall teaching performance in a course can be generalised from ratings from five or more classes taught by the instructor in which at least 15 students were enrolled in each class <sup>9</sup>.
4. The same instructor teaching different sections of the same course receives similar global ratings from each section <sup>10</sup>.

(Braskamp et al 1985; table 4.4 p42).

Overall, then, with larger classes, student ratings of instruction are reliable.

---

<sup>7</sup> Based on Crooks and Kane (1981); Feldman (1977; 1978); Marsh and Overall (1981); Marsh, Overall and Kessler (1979b).

<sup>8</sup> Based on Centra (1980).

<sup>9</sup> Based on Crooks and Kane (1981); Kane, Crooks and Gillmore (1976).

<sup>10</sup> Based on Shingles (1977); Overall and Marsh (1979).

## **VALIDITY OF STUDENT RATINGS**

Do students know a good lecturer, ie are student ratings actually measuring good teaching? Here validity means that the ratings are an accurate assessment of teaching quality, not other factors, like class size or personality of student.

McBean and Al-Nassri (1982) noted that "students strongly believed that student evaluations do measure teacher effectiveness ... while faculty only slightly agreed" (p278). This statement can be said to show face validity. Some would argue, though, that this is only valid as an indicator of student satisfaction.

Costin et al (1971), in an early review of the literature, suggest determining the validity of student ratings as a "match" between "students' subjective criteria" and "faculty members' goal in teaching" (p513). But the question is then, what is the basis on which students make their judgments? Consistently three items appear in studies that Costin et al review - knowledge, interest in subject, and preparation. However, this approach is difficult in practice, because other items are also important to students, and faculty and students disagree over the relative importance of each item.

So the approach to establishing validity has concentrated on criterion validity.

### **Objective Validation: Criterion Validation**

This concentrates on the relationship of ratings with other objective measures. The most common measure used is student learning (usually defined as the grade in the course examination).

In a now famous study in "Science", Rodin and Rodin (1972) found a negative correlation between the amount learned from classes, and their rating of the teacher. They used a subjective rating of the lecturer, and an objective measure of the amount of calculus learned. The conclusion of  $r = -0.75$  correlation threatened the validity of students' evaluation ratings.

But subsequent studies have consistently found positive correlations. Frey (1978) lists a number of problems with the Rodins study - for example, study based on teaching assistants rather than teachers who gave the main lectures. Further on in his article, after reviewing the studies since Rodins, Frey points out the need to study the "regular instructors", and to use "a rating form which emphasises the appropriate teaching traits" (p75). Marsh (1984) spends time to highlight methodological weaknesses with the Rodins study.



At the end of his meta-analysis of 41 studies, Cohen (1981) found a mean correlational index of 0.43 between student ratings and performance in examinations.

In another meta-analysis, McCallum (1984) examined 12 studies which used a global item evaluation of the instructor or course, and the correlation with student achievement. The average correlation was .064 for "course" and .101 for "instructor" (p155).

Doyle and Whitely (1974) used a beginners French course taught in 12 separate sections, with a common examination. There were significant correlations between level of specific ratings, and scores in the examination. When mean section ratings were used, the correlations were very small. The conclusion is that some items, but not all, are correlated to student learning.

Frey (1978) in testing the validity of the two dimensions of "skill" and "rapport", correlated each with examination scores. Using a course divided into multiple sections, taught by different instructors, but with a common syllabus, textbook, and examination. The median correlations are different: for the "skill" factor, it was  $r = 0.81$  but for "rapport" it was  $r = 0.29$ . "The two rating factors are clearly not the same in their ability to indicate which teachers were most effective in preparing their students for the final examination" (p87).

What is effective teaching measured in terms of student learning is an unresolved issue. Doyle (1975) feels that there is "a tendency for the instructors' expository clarity or presentation to relate to student learning as measured by fairly traditional course examinations" (p65).

Scriven (1981), however, states that "The best teaching is not that which produces the most learning, since what is learned may be worthless" (p248). The Instructional Development and Effectiveness Assessment (IDEA) (Hoyt 1973) treats student learning as the primary measure of teaching effectiveness, by including a section for the student to report their learning progress. Thus the criterion measure of effective teaching is part of the rating instrument.

Obviously this is open to criticism, but Cashin and Downey (1992) point out that "students who report learning more tend to score higher on an external examination" (p568), and there is support for validity of self-reports generally (eg: see Balk et al 1989).

Benton (1992), in a little known literature review of 31 studies correlating student achievement with ratings, is confident that "student evaluations of

instruction are tapping into an important dimension of teaching" (p40). But later admits that more research is needed as the significant correlations range from  $-.75$  to  $+.96$ .

Doyle (1983) lists his problems with using a student achievement test as the criterion for establishing the validity of student ratings of instruction:

- i) some characteristics of teaching are not linked to test scores - eg: "clarity" and "rapport";

- ii) it is assumed that the relationship is a linear one and thus the Pearson product-moment correlation can be used. But it is possible that it is a non-linear relationship between student achievement and student ratings of instruction;

- iii) which unit of analysis should be used:

- a) pooled within-class analysis (individual ratings in each section of the course, and average across course);

- b) between-sections analysis (mean ratings of evaluation items across course);

- c) total-class approach (individual ratings). Doyle prefers the first approach;

- iv) if subjects are randomly divided into sections of the course, then the generalisability of findings are limited.

The main alternative to final grade is to use students' gains in knowledge. But there are problems in how to measure the gain. Marsh and Overall (1980) tried to combine both criteria. They used final examination grade, ability to apply course material, and inclination to pursue the subject further. The first is seen as a cognitive criterion, while the other two are self-reported affective criteria. The students used were taking a course in computer programming. The authors, accepting methodological weaknesses, feel that more than one construct must be used to establish validity. "Therefore, because there is no universally accepted criterion of effective teaching, the validation of any teaching effectiveness measure must focus on a wide range of indicators" (p474).

Obviously, the higher the correlation, the better for validation. But validity will be specific to a particular situation, and "must always be evaluated in relation to a situation as similar as possible to the one

in which the measure is to be used" (Thorndike and Hagen 1977 p69).

## Construct Validation

For some researchers, criterion validity is not a satisfactory method to establish the validity of student ratings of instruction, because effective teaching is a construct. Thus for them construct validation is the best method. The main aim is to correlate multiple indicators of effective teaching. For example, student ratings and various criteria assessed for convergent and discriminant validity.

Howard et al (1985) use this method to establish teaching effectiveness using student ratings, colleagues ratings, teacher self-ratings, former-student ratings, and trained observers. Ratings by current and former students were most effective. Gaski (1987) is critical of this study.

A number of criteria are used under the heading of the Multi-Trait Multi-Method (MTMM) approach (Campbell and Fiske 1959). The use of a number of methods to measure one trait/construct allows correlations to be made; thus producing a MTMM matrix. It allows the estimation of variance due to traits or methods, and of unique or error variance.

It is possible to show convergent validity (correlation between items that should go together) and divergent validity (small or no correlation between items that should not go together). This method allows the research to estimate the effects of bias; for example, method bias: large correlation between variables because of the method used.

The main criteria used are self-evaluation by the lecturer, colleagues' evaluation, external observers, administrators, former students' evaluations, and the research productivity of lecturers.

### 1. LECTURER SELF-RATING.

There is a general tendency for instructors to rate themselves more favourably than their students do. But there is agreement on instructor's strengths and weaknesses. Centra (1972) found differences also between faculties: instructors in natural sciences rated effort needed for their course less than did the students, while education, business, home economics, and nursing instructors were the opposite.

Marsh (1982a), quoting his own studies, finds

correlations of  $r = 0.41$  for undergraduate ratings, and  $r = 0.39$  for postgraduate ratings, with lecturer's self-evaluation. Marsh (1984) is confident that this method demonstrates "acceptable validity", and also at undergraduate and postgraduate level.

Feldman (1989a) makes the comparison based on individual characteristics of teaching. Current students' and lecturers' self-evaluation are most similar in "stimulation of interest" and "availability and helpfulness", but less similar on "clarity of course objectives" and "intellectual expansiveness". Also lecturers rate themselves higher on "feedback", "friendliness", and "sensitivity" towards students.

## 2. RATINGS BY COLLEAGUES.

In their early literature review, Costin et al (1971) find correlations between 0.30 and 0.63 for students' ratings and colleagues' ratings. But in most cases, colleagues' ratings are not based on sitting through the lecture, but on "student hearsay, on the observation of the presumed effects of instruction ... and on inferences from their personal acquaintances (with the colleagues)" (Guthrie 1949 p113).

Ballard, Reardon and Nelson (1976) found correlations that range from 0.62 to 0.84. Studies based on colleagues actual visitation to the classroom are limited.

Furthermore, there is the problem that the presence of an observer can change the classroom situation - for example, by effecting the performance of the lecturer. Murray (1980) feels peer ratings are "less sensitive, reliable and valid" (p45) than student ratings.

## 3. OBSERVATION BY EXTERNAL OBSERVERS.

Murray (1980) feels that student ratings "can be accurately predicted from outside observer reports of specific classroom teaching behaviours" (p31). The feeling is that trained observers are best, and particularly if they concentrate on specific behaviour (eg: clarity-related behaviour: number of false starts or halts in speech, redundantly spoken words, and tangles in words) (Marsh 1984).

## 4. ADMINISTRATORS' VIEW.

Cotsonas and Kaiser (1962) used clinical students in a medical school, and compared their ratings with departmental administrators. The former tended to stress

the attitude towards students, and teaching skill, while the latter stressed knowledge. The authors suggest that the administrators noted the knowledge of the lecturer, and then assumed the other abilities ("halo effect"). It would also seem that the administrators took into account more than just classroom behaviour, but also their general judgments about the lecturer.

## 5. RETROSPECTIVE RATINGS OF ALUMNI.

Graduating students were asked to nominate "most outstanding" and "least outstanding" lecturers in their departments. Then undergraduates were asked to rate the nominated lecturers. Results indicated that the "most outstanding" lecturers were rated higher than the "least outstanding". A correlation of  $r = 0.82$  between graduates' and undergraduates' choices of most and least outstanding (Marsh 1977).

Gaski (1987) suggests caution when using former students' ratings for validity purposes because "the similarity between the student and former student teaching evaluations can be explained if the primary determinant of the former student ratings is former students' recollection of the assessment they made when they were current students of the given instructor one or two years earlier" (p329).

## 6. RESEARCH PRODUCTIVITY.

Blackburn (1974) suggested research and effective teaching were opposites. For example, McDaniel and Feldhusen (1970) found significant negative correlation between first authorship of books and students' ratings of teaching. But a significant positive correlation between second authorship of professional articles and rating of teaching.

Marsh (1984) finds no correlation or a small positive correlation between the two. "Although these findings seem to neither support nor refute the validity of student ratings, they do demonstrate that measures of research productivity cannot be used to infer teaching effectiveness or vice versa" (p729).

Feldman (1987), in another extensive review, looks at 43 studies of research productivity and overall teaching effectiveness, and finds a weak positive correlation. But when correlated with specific teaching abilities, there is a strong significant positive relationship with "knowledge of subject", and "preparation for classes".

## 7. OTHER CRITERIA.

Marsh (1987) briefly mentions other criteria for assessing the validity of students' ratings - enrolment in advanced courses of the same subject; instructor enjoyment of teaching; open-ended comments; whether students pursue the subject further (eg: Marsh and Overall 1980 computer students who rated lecturer highly were more likely to join local computer club).

Feldman (1989a) undertook a detailed literature review of the North American studies comparing overall ratings of teaching effectiveness made by current and former students, lecturers' colleagues, administrators, external (neutral) observers, and teachers' self-evaluation. The results are summarised in table 2.

Feldman concludes that there is similarity between various raters, in this order: current students and colleagues; current students and administrators; colleagues and administrators (similar in relative assessment, but not in absolute assessment); self-evaluation and current students; self-evaluation and colleagues. For the other relationships, there are not enough studies to determine.

| Method Used        | Current Students | Former Students | External Observers | Colleague | Administrators |
|--------------------|------------------|-----------------|--------------------|-----------|----------------|
| Current Students   |                  | +.69(6)*        | +.50(5)*           | +.55(14)* | +.39(11)*      |
| Former Students    |                  |                 | +.08(1)            | +.33(1)   | no cases       |
| External Observers |                  |                 |                    | -.12(1)   | no cases       |
| Colleague          |                  |                 |                    |           | +.48(5)*       |
| Administrators     |                  |                 |                    |           |                |

\* = significant correlation  $p < 0.001$  two-tailed. The number in ( ) is number of studies found.

Table 2 - showing a summary of the studies found by Feldman (1989a) showing a correlation between different methods of assessing teaching effectiveness.

The question of establishing validity has become a methodological issue debated in the literature, particularly around the use of criterion validity (established through multi-section courses) or construct validity (established using MTMM).

However, taking into account the weaknesses of the use of the different criteria, it is fair to say that

student ratings of instruction are valid. But the criteria used are validity measures of what?

Feldman (1977) looks at the purpose of the ratings - if it is to obtain objective descriptions of teachers, there may be a problem, but not if it is to measure students' subjective responses.

## **REFERENCES**

Abrami, P.C; Leventhal, L & Perry, R.P (1982) Educational seduction, Review of Educational Research, 52, 3, 446-464

Aleamoni, L.M & Thomas, G.S (1977) Is the instructor's rating of the class related to the class's rating of the instructor? Research Report No.1 O.I.R.D, Tucson, Arizona: Office of Instructional Research and Development, University of Arizona

Anikeef, A.M (1953) Factors affecting student evaluation of college faculty members, Journal of Applied Psychology, 37, 458-60

Arubayi, E.A (1987) Improvement of instruction and teacher effectiveness: are student ratings reliable and valid, Higher Education, 16, 267-278

Ballard, M; Reardon, J & Nelson, J (1976) Student and peer rating of faculty, Teaching of Psychology, 3, 88-90

Barke, C.R; Tollefson, N & Tracy, D.B (1981) Relationship between course entry attitudes and end-of-course ratings, Journal of Educational Psychology, 75, 1, 75-85

Bendig, A.W (1953) Student achievement in introductory psychology and student ratings of the competence and empathy of their instructors, Journal of Psychology, 36, 427-433

Benton, S.E (1992) Rating College Teaching: Criterion Validity Studies of Student Evaluation of Instructor Instruments, AAHE Research Report no.1; Washington: American Association for Higher Education

Blackburn, R.T (1974) The meaning of work in academia. In Doi, J (ed) Assessing Faculty Effort, San Francisco: Jossey Bass

Bradbury, P.S & Ramsden, P (1975) Student evaluation of teaching at North East London Polytechnic. In Evaluating Teaching in Higher Education, Collection of conference papers, London: University Teaching Management Unit

Braskamp, L; Brandenburg, D.C & Ory, J.C (1985) Evaluating Teaching Efficiency: A Practical Guide, Beverley Hills, CA: Sage

Brown, D.L (1976) Faculty ratings and student grades: a university-wide multiple regression analysis, Journal of Educational Psychology, 68, 5, 573-578

Bryant, P.T (1967) By their fruits ye shall know them, Journal of Higher Education, 38, 326-330

Byrne, D and Clore, G.L (1970) A reinforcement model of evaluation responses, Personality: An International Journal, 1, 103-128

Byrne, D & Nelson, D (1965) Attraction as a linear function of proportion of positive reinforcements, Journal of Personality and Social Psychology, 1, 659-663

Campbell, D.T & Fiske, D.W (1959) Convergent and discriminant validation by the MTMM matrix, Psychological Bulletin, 56, 81-105

Cashin, W.E & Downey, R.G (1992) Using global student rating items for summative evaluation, *Journal of Educational Psychology*, 84, 4, 563-572

Centra, J.A (1972) Two Studies on the Utility of Student Ratings for Improving Teaching, SIR Report no.2; Princeton, NJ: Educational Testing Service

Centra, J.A (1974) The relationship between student and alumni ratings of teachers, *Educational and Psychological Measurement*, 34, 321-325

Centra, J.A (1980) Determining Faculty Performance, San Francisco: Jossey Bass

Clark, K.E & Keller, R.J (1954) Student ratings of college teaching. In Eckert, R et al (eds) *A University Looks at Its Programme*, Minneapolis: University of Minneapolis

Cohen, P.A (1980) Effectiveness of student-rating feedback for improving college instruction: a meta-analysis, *Research in Higher Education*, 13, 321-341

Cohen, P.A (1981) Student ratings of instruction and student achievement: a meta-analysis of multi-section validity studies, *Review of Educational Research*, 51, 3, 281-309

Cooper, B & Foy, J (1967) Evaluating the effectiveness of lectures, *Universities Quarterly*, 21, 2, 182-185

Costin, F (1968) A graduate course in the teaching of psychology: description and evaluation, *Journal of Teacher Education*, 19, 425-432

Costin, F; Greenough, W.T & Menges, R.J (1971) Student ratings of college teaching: reliability, validity, and usefulness, *Review of Educational Research*, 41, 511-535

Cotsonas, N.J & Kaiser, H.F (1962) A factor analysis of students' and administrators' ratings of clinical teachers in a medical school, *Journal of Educational Psychology*, 53, 219-223

Crannell, C.W (1948) An experiment in the rating of instructors by their students, *College and University*, 24, 5-11

Crittenden, K.S & Norr, J.L (1973) Student values and teacher evaluation: a problem in person perception, *Sociometry*, 36, 2, 143-151

Crooks, T.J & Kane, M.J (1981) The generalisability of student ratings of instructors: item specificity and section effects, *Research in Higher Education*, 15, 305-313

Downie, N.W (1952) Student evaluation of faculty, *Journal of Higher Education*, 23, 495-496, 503

Doyle, K.O (1975) *Student Evaluation of Instruction*, Lexington, Mass: Lexington Books

Doyle, K.O (1983) *Evaluating Teaching*, Lexington, Mass: Lexington Books

Doyle, K.O & Whitely, S (1974) Student ratings as criteria of effective teaching, *American Educational Research Journal*, 11, 259-274

Drucker, A.J & Remmers, H.H (1951) Do alumni and students differ in their attitudes toward instructor? *Journal of Educational Psychology*, 42, 3, 129-143

Dunkin, M.J & Barnes, J (1986) Research on teaching in higher education. In Wittrock, M (ed) *Handbook of Research on Teaching*, London: Macmillan

Echandia, P.P (1964) A methodological study and factor analytic validation of forced choice performance ratings of college accounting instructors, *Dissertation Abstracts*, 25, 4, 2605-2606



Entwistle, N & Ramsden, P (1983) Understanding Student Learning, London: Croom Helm

Entwistle, N & Tait, H (1990) Approaches to learning, evaluations of teaching, and preferences for contrasting academic environments, Higher Education, 19, 169-194

Eysenck, H.J & Eysenck, S.B (1964) Manual of Eysenck Personality Inventory, London: University of London Press

Feldman, K.A (1976a) Grades and college students' evaluations of their courses and teachers, Research in Higher Education, 4, 69-111

Feldman, K.A (1976b) The superior college teacher from the students' view, Research in Higher Education, 5, 243-288

Feldman, K.A (1977) Consistency and variability among college students in rating their teachers and courses: a review and analysis, Research in Higher Education, 6, 223-274

Feldman, K.A (1978) Course characteristics and college students' ratings of their teachers: what we know and what we don't know, Research in Higher Education, 9, 199-242

Feldman, K.A (1979) The significance of circumstances for college students' ratings of their teachers and courses, Research in Higher Education, 10, 2, 149-172

Feldman, K.A (1983) Seniority and experience of college teachers as related to evaluations they receive from students, Research in Higher Education, 18, 1, 3-124

Feldman, K.A (1984) Class size and college students' evaluations of teachers and courses: a closer look, Research in Higher Education, 21, 1, 45-117

Feldman, K.A (1987) Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: a review and exploration, Research in Higher Education, 26, 3, 227-298

Feldman, K.A (1989a) Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers, Research in Higher Education, 30, 2, 137-194

Feldman, K.A (1992) College students' views of male and female college teachers: Part 1 - evidence from the social laboratory and experiments, Research in Higher Education, 33, 3, 317-375

Feldman, K.A (1993) College students' views of male and female college teachers: Part 2 - evidence from students' evaluations of their classroom teachers, Research in Higher Education, 34, 2, 151-212

Flood Page, C (1974) Student Evaluation of Teaching: The American Experience, London: Society for Research in Higher Education

Follman, J; Lucoff, M; Small, L & Power, F (1974) Kinds of keys to student ratings of faculty teaching effectiveness, Research in Higher Education, 2, 173-179

Foy, J (1969) A note on lecturer evaluation by students, Universities Quarterly, 23, 3, 345-349

Freehill, M.F (1967) Authoritarian bias and evaluation of college experiences, Improving College and University Teaching, 15, 18-19

Frey, P.W (1976) Validity of student instructional ratings: does timing matter? Journal of Higher Education, 67, 327-336

Frey, P.W (1978) A two-dimensional analysis of student ratings of instruction, Research in Higher Education, 9, 69-91

- Gage, N.L (1961) The appraisal of college teaching, *Journal of Higher Education*, 32, 17-22
- Gaski, J.F (1987) On 'construct validity of measures of college teaching effectiveness', *Journal of Educational Psychology*, 79, 3, 326-330
- Glass, G.V (1976) Primary, secondary and meta-analysis of research, *Educational Researcher*, 5, 3-8
- Glass, G.V (1978) Integrating findings: the meta-analysis of research. In Shulman, L.S (ed) *Review of Research in Education Vol 5*, Itasca, Illinois: F.E.Peacock
- Guthrie, E.R (1927) Measuring student opinion of teachers, *School and Society*, 25, 175-176
- Guthrie, E.R (1949) The evaluation of teaching, *Educational Record*, 30, 109-115
- Guthrie, E.R (1954) *The Evaluation of Teaching: A Progress Report*, Seattle: University of Washington
- Hildebrand, M & Wilson, R.C (1970) *Effective University Teaching and Its Evaluation*, Berkeley Centre for Research and Development in Higher Education: University of California
- Howard, G.S; Conway, G.C & Maxwell, S.E (1985) Construct validity measures of college teaching effectiveness, *Journal of Educational Psychology*, 77, 2, 187-196
- Hoyt, P.D (1973) Measurement of instructional effectiveness, *Research in Higher Education*, 1, 367-378
- Jones, J (1989) Students' ratings of teacher personality and teaching competence, *Higher Education*, 18, 551-558
- Jones, J; Lennie, M & Robinson, A (1985) Students' expectations of good teachers: primary, secondary and tertiary views, Paper presented at New Zealand Association of Research in Education, Auckland, New Zealand
- Kane, M.T; Crooks, T.J & Gillmore, G.M (1976) Generalisability and the Interpretation of Student Evaluation of Teachers, Paper presented at Annual Meeting of American Educational Research Association, San Francisco
- Kappes, M.M (1988) *A Comparison of Students' Ratings of Full-time and Part-time Instructors' Teaching Effectiveness in a Community College*, Unpublished PHD Thesis, University of Pittsburgh
- Kirchner, R.P (1967) *A Central Factor in Teacher Evaluation by Students*, Unpublished research paper, College of Education: Lexington University of Kentucky
- Kirker, M.J (1990) *Variance in Student Ratings of Part-time and Full-time Instructor Effectiveness by Teaching Field and Function of at a Mid-Western Community College*, Unpublished PHD Thesis, Iowa State University
- Klyczek, J.P (1989) *A Study of faculty Characteristics Affecting Student Evaluation of Instruction in Higher Education*, Unpublished PHD Thesis, State University of New York at Buffalo
- McBean, E.A & Al-Nassri, S (1982) Questionnaire design for student measurement of teaching effectiveness, *Higher Education*, 11, 273-288
- McCallum, L.W (1984) A meta-analysis of course evaluation data and its use in the tenure decision, *Research in Higher Education*, 21, 2, 150-157
- McClelland, J.N (1970) The effect of student evaluations of college instruction upon subsequent evaluations, *California Journal of Educational Research*, 21, 88-95
- McDaniel, E.D & Feldhusen, J.F (1970) *Relationships between faculty*

ratings and indexes of service and scholarship, Proceedings of the 78th Annual Convention of the American Psychological Association, 5, 619-620

McKnight, P (1973) Curriculum and Instructional Survey, University of Kansas: Office of Instructional Resources

Marsh, H.W (1977) The validity of students' evaluations: classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors, American Educational Research Journal, 14, 4, 441-447

Marsh, H.W (1982a) SEEQ: a reliable, valid and useful instrument for collecting students' evaluations of university teaching, British Journal of Educational Psychology, 52, 77-95

Marsh, H.W (1982b) Factors affecting students' evaluations of the same course taught by the same instructor on different occasions, American Educational Research Journal, 19, 485-497

Marsh, H.W (1984) Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases, and utility, Journal of Educational Psychology, 76, 5, 707-754

Marsh, H.W (1985) Student ratings of teaching. In Husen, T & Postlethwaite, T.N (eds) International Encyclopaedia of Education: Research and Studies Vol 8, Oxford: Pergamon

Marsh, H.W (1987) Students' evaluations of university teaching: research findings, methodological issues, and directions to future research, International Journal of Educational Research, 11, 253-388

Marsh, H.W & Cooper, T.L (1981) Prior subject interest, students' evaluations, and instructional effectiveness, Multivariate Behavioural Research, 16, 83-104

Marsh, H.W & Overall, J.U (1980) Validity of students' evaluations of teaching effectiveness: cognitive and affective criteria, Journal of Educational Psychology, 72, 4, 468-475

Marsh, H.W & Overall, J.U (1981) The relative influence of course level, course type and instructor on students' evaluations of college teaching, American Educational Research Journal, 18, 1, 103-112

Marsh, H.W; Overall, J.U & Kesler, S.P (1979b) Class size, students' evaluations and instructional effectiveness, American Educational Research Journal, 16, 1, 57-70

Martin, E (1984) Power and authority in the classroom: sexist stereotyping in teaching evaluations, Journal of Women in Culture and Society, 9, 482-492

Miklich, D.R (1969) An experimental validation study of the Purdue Rating Scale for Instruction, Educational and Psychological Measurement, 29, 963-967

Miller, R.I (1972) Evaluating Faculty Performance, San Francisco: Jossey-Bass

Murray, H.G (1980) Evaluating University Teaching: A Review of Research, Toronto: Ontario Confederation of University Faculty Associations

Naftulin, D.H; Ware, J.E & Donnelly, F.A (1973) The Doctor Fox lecture: a paradigm of educational seduction, Journal of Medical Education, 48, 630-635

Overall, J.U & Marsh, H.W (1979) Midterm feedback from students: its relationship to instructional improvement and students' cognitive and affective outcomes, Journal of Educational Psychology, 71, 856-865

Overall, J.U & Marsh, H.W (1980) Students' evaluation of instruction: a longitudinal study of their stability, Journal of Educational Psychology,

Pohlmann, J.T (1972) Summary of Research on the Relationship Between Student Characteristics and Student Evaluations of Instruction at Southern Illinois University, Carbondale, Technical Report 1.1 - 72, Carbondale, Ill: Counselling and Testing Centre, Southern Illinois University

Prosser, M & Trigwell, K (1990) Student evaluations of teaching and courses: student study strategies as a criterion of validity, Higher Education, 20, 135-142

Rees, R.D (1969) Dimensions of students' point of view in rating college teachers, Journal of Educational Psychology, 60, 476-482

Remmers, H.H(1960) Manual of Instruction for the Purdue Rating Scale for Instructors (rev ed), West Lafayette, Ind: University Book Store

Remmers, H.H; Martin, F.D & Elliot, D.N(1949) Are students' ratings of instructors related to their grades? Purdue University Studies in Higher Education, 66, 17-26

Rezler, A.G (1965) The influence of needs upon the student's perception of his instructor, Journal of Educational Research, 58, 282-286

Riley, J.W; Ryan, B.F & Lifschitz, M (1950) The Student Looks at his Teacher, New Brunswick, MJ: Rutgers University Press

Rodin, M & Rodin, B (1972) Student evaluations of teachers, Science, 177, 1164-1166

Rokeach, M (1960) The Open and Closed Mind, New York: Basic Books

Rubinstein, J & Mitchell, H (1970) Feeling free, student involvement and appreciation, Proceedings of the 78th Annual Convention of the APA, 623-624

Scriven, M (1981) Summative teacher evaluation. In Millman, J (ed) Handbook of Teacher Evaluation, Beverley Hills, CA: Sage

Sharon, A.T & Bartlett, C.J (1969) Effects of instructional conditions in producing leniency on two types of rating scales, Personnel Psychology, 22, 251-263

Shingles, R.D (1977) Faculty ratings: procedures for interpreting student evaluations, American Educational Research Journal, 14, 459-470

Smith, B.G.N; Fusilier, C.N; Bagramian, R.A & Bottomley, W.K (1969) A criterion model of the dental school instructor, Journal of Dental Education, 33, 523-531

Smithers, A (1970b) What do students expect of lectures? Universities Quarterly, 24, 330-336

Tetenbaum, T (1977) The factor invariance of student ratings of instruction under three sets of directions, Research in Higher Education, 6, 11-23

Thorndike, R.L & Hagen, E (1977) Measurement and Evaluation in Psychology and Education (4th ed), New York: John Wiley

Tollefson, N; Chen, J.S & Kleinsasser, A (1989) The relationship of students' attitudes about effective teaching to students' ratings of effective teaching, Educational and Psychological Measurement, 49, 529-536

Whitten, B.J & Umble, M.M (1980) The relationship of class size, class level, core vs non-core classification for a class to student ratings of faculty: implications for validity, Educational and Psychological Measurement, 40, 419-423

Wiggington, H; Tollefson, N & Rodriguez, E (1989) Students' ratings of

instructors revisited: interactions among class and instructor variables, Research in Higher Education, 30, 3, 331-344

Yonge, G.D & Sassenrath, J.M (1968) Student personality correlates of teacher ratings, Journal of Educational Psychology, 59, 44-52

## **APPENDIX 1**

### **OTHER LESS IMPORTANT POTENTIAL BIASES OF STUDENT RATINGS OF INSTRUCTION**

#### **1. STYLE OF LEARNING.**

Entwistle and Ramsden (1983) proposed 4 styles or approaches to learning:

- a) Deep approach - students attempt to understand rather than just accept, using other approaches.
- b) Comprehension learning - building overall description of content and link to previous knowledge.
- c) Operation learning - detailed attention to evidence.
- d) Surface approach - memorization.

Students were allocated to a style of learning by Lancaster Approaches to Study Inventory, then given the Course Perceptions Questionnaire. The general conclusion, which was replicated seven years later is "that students who adopt meaning or reproducing orientations also prefer the methods of teaching and assessing which encourage those approaches to learning" (Entwistle and Tait 1990 p188).

Confirmed by Prosser and Trigwell (1990) in Australia: "courses in which students adopted deeper approaches to study were also the courses that had teaching that was rated more highly" (p 141).

#### **2. TEACHER PERSONALITY.**

Jones (1989) tried to investigate what students actually evaluate about the instructor - is it really the course/teaching, or their personality? After analysis of the results, it was found that the student ratings of teacher personality loaded on the same factor as their rating of teacher competence. Thus teacher personality is seen as part of teaching competence, and that "in fact it would be very surprising if students' perception of a teacher's personality did not affect their rating of her or his teaching competence" (p556).

Jones et al (1985) report that students look at two aspects of the teacher: technical aspects (ability to explain/knowledge of subject) and personological aspects (personality - eg: listens to students).

Crittenden and Norr (1983) tried to apply a "person perception" model to teacher evaluation, and sees it as a special case of person perception.

Flood Page (1974) says the relationship is "to say the least, obscure" (p55). It is not always easy "to separate best from worst teachers on personality grounds" (p52). Costin et al (1971) agree after reviewing 12 studies.

Furthermore, mere popularity is not enough. But Guthrie (1954) did find that students rated higher those instructors, who had great interest/enthusiasm for their subject.

Nor is there any relationship between the teacher's activities outside the classes (ie: allocation of time to research/preparation etc), and good/bad teaching (Hildebrand and Wilson 1970).

### 3. STUDENT'S SENIORITY.

This can be looked at as actual age of student or year of course. Studies vary from finding that senior students rate higher (eg: Whitten and Umble 1980) to no relationship (eg: Marsh and Overall 1981).

Smith et al (1969) showed that students' attitudes to what is good teaching on a dental course changes over time, particularly on three items: "is cognizant of student problems"; "encourages student judgment"; and "possesses current knowledge of subject". But the general ratings did not change (quoted in Flood Page 1974).

### 4. SIMILARITY BETWEEN TEACHER/STUDENT.

Tollefson et al (1989) looked at the question of whether students would rate higher a teacher who held the same attitudes to themselves about what is effective teaching. Based on the social psychological theory that individuals are attracted to persons who hold similar views (Byrne and Clore 1970; Byrne and Nelson 1965). Earlier studies were unclear. Tollefson et al used the Attitude Toward Effective Teaching Scale (ATET), and the Teacher Rating Scale (TRS)(McKnight 1973). This study was also inconclusive - two separate analyses produced conflicting results.

Feldman (1977) "There are hints that under some circumstances similarity of teacher-student gender is

associated with higher ratings" (p245).

## 5. FORMAL TRAINING.

Costin (1968) noted that General Teaching Assistants (GTA) in psychology who attended a short teaching course received higher ratings in "feedback" and "group interaction" than those who had not.

## 6. IMPRESSIONS OF INSTRUCTOR.

Overall impression:

There is evidence that the overall impression of instruction can influence specific ratings of a lecturer. For example, Pohlmann (1972) found a correlation of approx 0.2 between overall evaluation of instruction at Southern Illinois University, and specific teacher ratings. Other studies find varying correlations.

Initial impression:

Feldman (1977) quotes studies suggesting that between one fifth to one third of variance in final ratings is due to the students' early impressions.

Pre-course impressions:

Students who have heard a professor is good, rate them higher than those who have not heard about the professor (Miller 1972). But there is a selection effect here - students are more likely to select courses taught by instructors they have heard good comments about or have had good experiences with before, than unfamiliar instructors, or ones who have received poor reports.

However, there is concern over the effects of pre-course expectations. Barke et al (1983) compared responses on the Affective Entry Questionnaire and Course Evaluation Questionnaire. Respondents tend to answer "no basis on which to make judgment" in the first questionnaire, suggesting "that, as a rule, students may have fewer expectations or biases that could potentially influence end-of-course ratings than many instructors believe" (p83).

## 7. STUDENT ABILITY.

No relationship of either type. Remmers et al (1949) explain these results by the fact that teaching is aimed at the whole class. For example, it may be too slow for brighter students, leading to a poorer rating of the teacher, but just right for slower students leading to a favourable rating.

## 8. MISCELLANEOUS FACTORS.

Arubayi (1987) adds time of day (morning lectures rated higher), and mood of students. Other evidence on time of day inconsistent (see Feldman 1979 p219).

McClelland (1970) divided students into 3 groups randomly: normal ratings forms given to one group; rating forms that contained alleged previous ratings, but artificially high to another group; and the same to the last group, but ratings artificially low. Significant differences found for groups 2 and 3; ie: higher or lower ratings respectively. Student ratings can, thus, be easily influenced it was suggested.

Students' feeling of control significantly correlated to appreciation of instructor (Rubinstein and Mitchell 1970).

The fear that students who are hostile to the lecturer may give them poor ratings is not borne out by Crannell (1948). However, there is little other research.

Kappes (1988) compared ratings of full-time and part-time lecturers. The latter rated significantly higher on "treating students with respect" and "starting/ending class on time". Full-timers rated higher on 8 items. This was confirmed by Kirker (1990).

Doyle (1982) suggests that based on common sense, events outside the classroom could influence the evaluation - eg: the day before a big event, or the busy last week of term.

## INTERACTION OF BIASES

Wigington et al (1989) looked at the interaction between class type (ie: lecture or seminar etc); class level; class size; instructor reputation, rank and sex. The data were analysed through 15 two-way factorial analyses of variance. The interactions found are detailed in the table 3.



| INTERACTION/VARIABLES                          | EFFECT OF INTERACTION                                                                                           |
|------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| Type of course by size of class                | Lecture-discussion: small classes rated lower. Lab format: small/large classes rated lower than medium sized.   |
| Level of course by sex of instructor           | Male teachers rated higher on higher courses.                                                                   |
| Type of course by rank of instructor           | No consistent pattern.                                                                                          |
| Rank of instructor by size of class            | Teaching assistants have U shaped profile, professors negative correlation.                                     |
| Level of course by rank of instructor          | Associate professors have highest rating from highest course.                                                   |
| Sex of instructor by size of class             | Male teachers higher rating on larger class.                                                                    |
| Rank of instructor by sex of instructor        | Professors higher ratings if females.                                                                           |
| Type of course by sex of instructor            | Female teachers higher rating for lecture, discussion and lab classes, but lower for lecture-discussion format. |
| Type of course by level of course              | Postgraduate courses rated lower for lectures and lab.                                                          |
| Reputation of instructor by type of course     | Higher rating for lecture and lab formats by teacher with reputation.                                           |
| Level of course by size of class               | Higher courses moderate-sized classes lower rating than large classes.                                          |
| Reputation of instructor by rank of instructor | Professors highest rating when reputation important, and lowest when not.                                       |

Table 3 - showing interaction of variables producing significant relationships in study by Wigington et al (1989).

There was no significant relationship for reputation by level, reputation by sex, and reputation by size. The authors conclude that " student ratings do reflect differences in instructional effectiveness". But "an interpretation of student ratings needs to reflect an understanding of the variables that interact to produce differences in student ratings of instructors" (p342).

Klyczek (1989) developed a path analysis of professional rank, age, gender, status, communication skills, relationship with students, availability to students, and publishing productivity. All variables had stronger relationships with each other than to student ratings of instruction.

## **APPENDIX 2**

### **CLASS SIZE AND INDIVIDUAL CHARACTERISTICS OF TEACHING EFFECTIVENESS**

| CHARACTERISTIC OF TEACHING | POSITIVE CORR-ELATION | NEGATIVE CORR-ELATION | NO CORR-ELATION | OTHER RELATION-SHIP |
|----------------------------|-----------------------|-----------------------|-----------------|---------------------|
| 1.Stimulation              | 1                     | 8                     | 12              | 1                   |
| 2.Enthusiasm               | 1                     | 3                     | 7               | 1                   |
| 3.Knowledge                | 1                     | 3                     | 5               | 0                   |
| 4.Expansive                | 0                     | 4                     | 4               | 1                   |
| 5.Preparation              | 3                     | 11                    | 12              | 3                   |
| 6.Clarity                  | 1                     | 9                     | 14              | 2                   |
| 7.Elocution                | 0                     | 5                     | 3               | 1                   |
| 8.Sensitivity              | 0                     | 6                     | 1               | 1                   |
| 9.Objectives               | 0                     | 5                     | 8               | 2                   |
| 10.Materials               | 0                     | 11                    | 11              | 3                   |
| 11.Materials               | 0                     | 5                     | 3               | 1                   |
| 12.Outcome                 | 2                     | 8                     | 3               | 2                   |
| 13.Fairness                | 1                     | 17                    | 8               | 2                   |
| 14.Personality             | 2                     | 1                     | 0               | 0                   |
| 15.Feedback                | 0                     | 7                     | 3               | 1                   |
| 16.Questions               | 1                     | 19                    | 5               | 3                   |
| 17.Challenge               | 1                     | 11                    | 4               | 2                   |
| 18.Respect                 | 1                     | 14                    | 2               | 3                   |
| 19.Availability            | 0                     | 15                    | 5               | 3                   |

Table 4 - summarising the number of studies found by Feldman (1984), showing the relationship between class size and different characteristics of teaching.

| Individual Characteristics of Ideal Lecturer | Relationship between individual characteristic and class size |
|----------------------------------------------|---------------------------------------------------------------|
| 1.Stimulation of interest                    | No correlation.                                               |
| 2.Enthusiasm                                 | No correlation.                                               |
| 3.Knowledge of subject                       | No correlation.                                               |
| 4.Intelligence                               | No correlation.                                               |
| 5.Preparation/organisation                   | Negative correlation.                                         |
| 6.Clarity                                    | No correlation.                                               |
| 7.Elocutionary skills                        | Negative correlation.                                         |
| 8.Class level                                | Negative correlation.                                         |
| 9.Course objectives                          | No correlation.                                               |
| 10.Practical                                 | Negative correlation.                                         |
| 11.Use of aids                               | Negative correlation.                                         |
| 12.Perceived outcome                         | Negative correlation.                                         |
| 13.Fairness                                  | Negative correlation.                                         |
| 14.Personality                               | Positive correlation.                                         |
| 15.Feedback                                  | Negative correlation.                                         |
| 16.Encourages questions                      | Negative correlation.                                         |
| 17.Encourage independent thought             | Negative correlation.                                         |
| 18.Respect                                   | Negative correlation.                                         |
| 19.Availability                              | Negative correlation.                                         |

Table 5 - showing the most common relationship between class size and individual characteristics of teaching, as found by Feldman (1984).

## **APPENDIX 3**

### **CORRELATION OF INDIVIDUAL ITEMS WITH OVERALL EVALUATION OF TEACHING EFFECTIVENESS**

| Individual Characteristics of Ideal Lecturer | Correlation of individual characteristic with overall evaluation |
|----------------------------------------------|------------------------------------------------------------------|
| 1.Stimulation of interest                    | +.20                                                             |
| 2.Enthusiasm                                 | +.46                                                             |
| 3.Knowledge of subject                       | +.48                                                             |
| 4.Intelligence                               | +.54                                                             |
| 5.Preparation/organisation                   | +.41                                                             |
| 6.Clarity                                    | +.25                                                             |
| 7.Elocutionary skills                        | +.49                                                             |
| 8.Class level                                | +.40                                                             |
| 9.Course objectives                          | +.45                                                             |
| 10.Practical                                 | +.70                                                             |
| 11.Use of aids                               | +.72                                                             |
| 12.Perceived outcome                         | +.28                                                             |
| 13.Fairness                                  | +.72                                                             |
| 14.Personality                               | -                                                                |
| 15.Feedback                                  | +.87                                                             |
| 16.Encourages questions                      | +.60                                                             |
| 17.Encourage independent thought             | +.39                                                             |
| 18.Respect                                   | +.65                                                             |
| 19.Availability                              | +.74                                                             |

Table 6 - showing the correlation with overall evaluation of individual characteristics of teaching, as found by Feldman (1976b).